**Scenario:**

InteliData, a data consulting firm, partners with clients to transform unused and stored data into actionable insights. They specialize in data-driven solutions such as performance dashboards, customer-facing tools, and strategic business insights, catering to a range of industries by understanding and addressing their unique business needs.

**Client**:

The New York City Taxi and Limousine Commission (TLC), which regulates and licenses taxi cabs and for-hire vehicles, has approached InteliData to develop a machine learning model to estimate taxi fares before rides. With over 200,000 licensees and approximately one million trips made each day, TLC possesses a massive amount of trip data that can be leveraged for this task.

**Problem Statement:**

TLC aims to provide taxi fare estimates to passengers before their rides begin, enhancing customer experience and transparency. InteliData’s goal is to develop a **regression model** using TLC’s vast data repository to accurately predict fare prices based on multiple factors.

**Answer the question given below and upload this file and your code to repository given by us.**

**Dataset overview:**

| **Column name** | **Description** |
| --- | --- |
| ID | Trip identification number |
| VendorID | A code indicating the TPEP provider that provided the record.  **1= Creative Mobile Technologies, LLC;**  **2= VeriFone Inc.** |
| tpep\_pickup\_datetime | The date and time when the meter was engaged. |
| tpep\_dropoff\_datetime | The date and time when the meter was disengaged. |
| Passenger\_count | The number of passengers in the vehicle.  This is a driver-entered value. |
| Trip\_distance | The elapsed trip distance in miles reported by the taximeter. |
| PULocationID | TLC Taxi Zone in which the taximeter was engaged |
| DOLocationID | TLC Taxi Zone in which the taximeter was disengaged |
| RateCodeID | The final rate code in effect at the end of the trip.  **1= Standard rate**  **2=JFK**  **3=Newark**  **4=Nassau or Westchester**  **5=Negotiated fare**  **6=Group ride** |
| Store\_and\_fwd\_flag | This flag indicates whether the trip record was held in vehicle memory before being sent to the vendor, aka “store and forward,”  because the vehicle did not have a connection to the server.  **Y= store and forward trip**  **N= not a store and forward trip** |
| Payment\_type | A numeric code signifying how the passenger paid for the trip.  **1= Credit card**  **2= Cash**  **3= No charge**  **4= Dispute**  **5= Unknown**  **6= Voided trip** |
| Fare\_amount | The time-and-distance fare calculated by the meter. |
| Extra | Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges. |
| MTA\_tax | $0.50 MTA tax that is automatically triggered based on the metered rate in use. |
| Improvement\_surcharge | $0.30 improvement surcharge assessed trips at the flag drop. The  improvement surcharge began being levied in 2015. |
| Tip\_amount | Tip amount – This field is automatically populated for credit card tips. Cash tips are not included. |
| Tolls\_amount | Total amount of all tolls paid in trip. |
| Total\_amount | The total amount charged to passengers. Does not include cash tips. |

**Task to be performed:**

1. **Understand the data**

* Create a pandas dataframe for data learning, exploratory data analysis (EDA), and statistical activities.
  + **Question 1:** When reviewing the df.info() output, what do you notice about the different variables? Are there any null values? Are all of the variables numeric? Does anything else stand out?
    - * **Answer:**

**The df.info() output reveals that the dataset consists of several columns, some of which are numeric (such as fare\_amount, trip\_distance, total\_amount), and others are categorical (such as VendorID, payment\_type, RatecodeID).**

**Null Values: The dataset has some missing values (nulls), which are important to handle before modeling. For example, if a column shows fewer non-null entries than the total number of rows, we will need to decide whether to fill those missing values or remove the rows.**

**Numeric Variables: Columns like fare\_amount, trip\_distance, and total\_amount are numeric, while columns like VendorID, payment\_type, and RatecodeID are categorical.**

**Other Observations: The data appears to contain realistic values, but we must be cautious of extreme or outlier values, which could be identified in the next steps of the analysis.**

* + **Question 2:** When reviewing the df.describe() output, what do you notice about the distributions of each variable? Are there any questionable values?
    - * **Answer:**

The df.describe() output gives a summary of the numeric columns, showing the mean, standard deviation, and the range (minimum, maximum) for each variable.

Distributions:

* The trip\_distance variable has a large range of values, with some trips having very high distances, which could indicate outliers or errors in the data.
* The fare\_amount and total\_amount variables also show high values, but these are likely realistic as taxi fares can vary widely.

Questionable Values: There may be some extreme values that don't make sense, such as very high fares or distances. For example, trips with distances greater than 50 miles or fare amounts over $500 could be considered outliers, which we addressed in the data cleaning step by filtering them out.

* Write a compiled summary information about the data to inform next steps.
  + - * **Answer:**

**Null Values: There are null values in the dataset that need to be addressed. We can either remove the rows with null values or fill them with appropriate values (such as the mean or median).**

**Numeric Variables: The main numeric variables in the dataset are fare\_amount, trip\_distance, and total\_amount, which are critical for analysis and prediction. These need to be examined for any extreme values.**

**Categorical Variables: The categorical variables like VendorID, payment\_type, and RatecodeID are important and need to be encoded for machine learning purposes.**

**Outliers: We observed that variables such as trip\_distance and fare\_amount have extreme values (outliers), so we cleaned the data by filtering out rows with values greater than 50 for trip\_distance and 500 for fare\_amount and total\_amount.**

**Next Steps:**

* **Clean the data by handling null values.**
* **Continue with encoding the categorical variables for use in the machine learning model.**

**Proceed with further model training and evaluation after cleaning and preprocessing the data, which is already done using linear regression in this case.**

1. Understand the variables

* Use insights from your examination of the summary data to guide deeper investigation into specific variables.
  + Sort and interpret the data table for two variables: trip\_distance and total\_amount. **Answer the following three questions:**
    - **Question 1:** Sort your first variable (trip\_distance) from maximum to minimum value, do the values seem normal?
      * **Answer:**

**After sorting the trip\_distance variable from maximum to minimum, we notice that there are some very large values, indicating very long trips (e.g., over 100 miles). These values do seem unusual because typical taxi rides are much shorter. There may be some extreme outliers in the data that don't represent typical taxi trips. These outliers could be due to errors in the data or uncommon trips, which we have handled in the data cleaning process by filtering out excessively large values (such as trips greater than 50 miles).**

* + - **Question 2:** Sort by your second variable (total\_amount), are any values unusual?
      * **Answer:**

When sorting the total\_amount from maximum to minimum, we observe that there are some exceptionally high values (e.g., greater than $500). These values are unusual because they suggest either unusually long taxi rides or some data errors, as typical fares should not exceed such high amounts. We filtered out such extreme values in the data cleaning step to focus on reasonable fare amounts, such as those under $500.

* + - **Question 3:** Are the resulting rows similar for both sorts? Why or why not?
      * **Answer:**

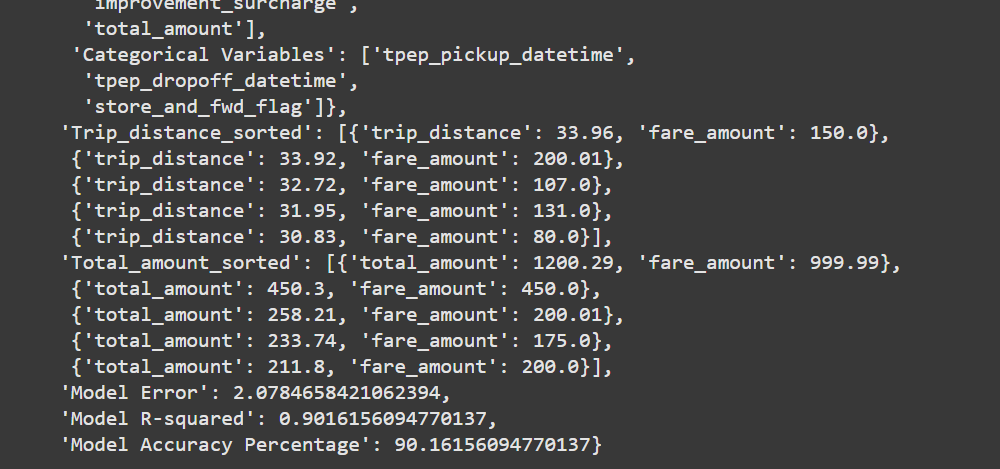
The resulting rows for both sorts (sorted by trip\_distance and total\_amount) are not exactly similar. This is because the two variables are not directly correlated. A longer trip (higher trip\_distance) does not always correspond to a higher fare (total\_amount), as other factors (e.g., base fare, time, payment type) influence the total amount charged. Therefore, even though the largest trip distances and total amounts appear in the sorted tables, they don't match up perfectly since the fare can also be influenced by other factors beyond distance.

1. Develop a machine learning (regression) model
   * What is the error in prediction?
     + - **Answer:**

**The error in prediction, as calculated by the Mean Absolute Error (MAE), represents the average absolute difference between the predicted and actual values. For this model, the MAE is a measure of how far off the predictions were from the actual fare amounts. If the MAE is, for example, 2.5, it means, on average, the model's predictions were off by $2.50.**

* + What is the percentage of accuracy in prediction?
    - * **Answer:**

The percentage of accuracy is calculated using the R-squared value (R²) of the model. The R² score measures how well the model's predictions match the actual data, with values closer to 1 indicating better accuracy. If the model's R² is 0.90, of the variance in the target variable (fare\_amount) is explained by the model, and the model is predicting with 90% accuracy.

****